

Finding Statistically Significant Attribute Interactions

Andreas Henelius*

Antti Ukkonen*

Kai Puolamäki*

Abstract

In many data exploration tasks it is meaningful to identify groups of attribute interactions that are specific to a variable of interest. These interactions are also useful in several practical applications, for example, to gain insight into the structure of the data, in feature selection, and in data anonymisation. We present a novel method, based on statistical significance testing, that can be used to verify if the data set has been created by a given factorized class-conditional joint distribution, where the distribution is parametrized by partition of its attributes. Furthermore, we provide a method, named ASTRID, to automatically find a partition of attributes that describes the distribution that has generated the data. The state-of-the-art classifiers are utilized to capture the interactions present in the data by systematically breaking attribute interactions and observing the effect of this breaking on classifier performance. We empirically demonstrate the utility of the proposed method with real and synthetic data as well as with usage examples.

Keywords: attribute interactions, constrained randomisation, hypothesis testing, clustering

1 Introduction

It is often of interest to understand the attribute interactions that are specific to a variable of interest. In this paper we consider interactions in the context of supervised learning. We say that two or more attributes are interacting if they carry complementary information and are jointly needed for predicting the class of a data item. Knowledge of the attribute interactions specific to a variable of interest has several important real-world applications, e.g., in feature selection and data anonymisation as we show in this paper.

However, finding interacting attributes in a dataset in the general case is not straightforward and requires complex modeling. In this paper we instead focus on leveraging classifiers to find interacting attributes, similarly to [8]. We may assume that state-of-the-art high-performing classifiers must at least implicitly model and utilise these complex attribute interactions, if they are able to make accurate predictions. This means, that if we can observe which attributes a classifier is jointly using when predicting class labels, we are able to deduce which attributes are interacting.

In this paper we focus on developing a novel method for finding a disjoint partition (grouping) $\mathcal{S} =$

$\{S_1, \dots, S_k\}$ of the attributes of a dataset, such that attributes in the same group S_i are interacting (dependent) given the class, and attributes in different groups are independent given the class, respectively.

Given a data matrix X , where the rows correspond to data items and columns attributes, respectively, and an associated vector of class labels C , a classifier tries to model the class probabilities given the data, i.e., to find $P(C | X) \propto P(X | C)P(C)$. Here $P(X | C)$ is the *class-conditional* distribution of the attributes. In this paper we focus on this distribution.

The grouping \mathcal{S} represents a factorisation of $P(X | C)$ into independent factors, i.e., $P(X | C; \mathcal{S}) = \prod_{S \in \mathcal{S}} P(X_S | C)$, where X_S only contains the attributes in the set S . In this paper an attribute interaction means that interacting attributes must be in the same group in \mathcal{S} and, hence, in the same factor in $P(X | C; \mathcal{S})$. Therefore, the attributes in a group are needed jointly to provide evidence regarding the class C .

We approach this problem of finding a factorisation using a principled statistical significance testing methodology. Our method is based on the following intuition. Assume that we train a classifier f_1 using data from a unfactorised distribution, and that we train a classifier f_2 using data from a factorised distribution. Now, if the classifiers f_1 and f_2 cannot be distinguished from each other in terms of performance, it means that the factorisation correctly captures the class-dependent structure in the data. On the other hand, if f_2 performs worse than f_1 it means that some essential relationships in the data needed by the classifier are no longer present, i.e., the factorisation is invalid.

In this paper we consider the two problems of (i) testing whether an attribute grouping \mathcal{S} correctly captures the attribute interaction structure, and (ii) automatically finding the maximum-cardinality attribute grouping \mathcal{S} for a dataset, corresponding to the maximally factorised joint class-conditional distribution. The first problem is solved using a randomisation test and for solving the second problem we present the novel ASTRID method.

By solving these problems we gain insight into *the structure of the data* and this has important applications in many domains. We next consider a few ex-

*Finnish Institute of Occupational Health, PO Box 40, FI-00251 Helsinki, Finland. Email: firstname.lastname@ttl.fi

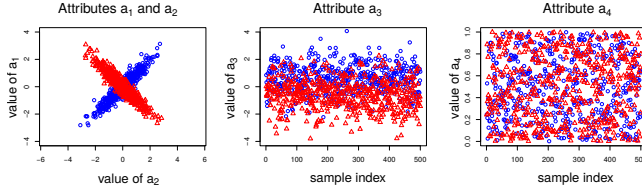


Figure 1: The synthetic dataset used as a running example. Class 0: blue circles, class 1: red triangles.

amples demonstrating the practical utility of attribute interactions.

Running Example As a running example we use a *synthetic dataset* D with 4 attributes a_1, \dots, a_4 and class labels C . The dataset is visualised in Fig. 1. It has two classes, each with 500 samples. The dataset is constructed so that attributes a_1 and a_2 carry meaningful class information only when considered jointly. Attribute a_3 contains some class information whereas attribute a_4 is random noise. The class-conditional attribute interaction structure is hence given by the grouping $\mathcal{S} = \{\{1, 2\}, \{3\}, \{4\}\}$, having three groups. Attributes in different groups in \mathcal{S} are independent and constitute the factors of P . We next exemplify the importance of attribute interactions and their wide applicability by discussing real-world examples.

Example 1. Attribute interactions have been used, e.g., in medicine to find attribute combinations that together constitute risk factors for a procedure [12]. In pharmacovigilance the interactions are important to understand which drug combinations that can cause adverse drug reactions [9]. In these cases we want to *identify groups of interacting attributes*, e.g., the grouping \mathcal{S} for the running example. In Sec. 6 we show attribute interactions in some datasets.

Example 2. In several data analysis applications we need to sample data such that some aspect of the data remains intact, while the data is otherwise random. E.g., our goal can be to shuffle a dataset such that the ability of a classifier to accurately make predictions from the data remains approximately intact. This has applications in, e.g, creating synthetic datasets for use in model compression [4], or data anonymisation [1, 3, 19] exemplified in Sec. 6. As shown in this paper, knowing the attribute interactions allows us to break the attribute interactions in the data not used by the classifier while the class-conditional joint distribution remains essentially the same. The important implication of this is that the classification performance on such a randomised (e.g., anonymised) dataset also remains

essentially unchanged. We here show how this can be done with statistical guarantees.

Example 3. An important problem in the analysis of large datasets is *variable and feature selection* to reduce training time for models, reduce the amount of storage space or to improve classification performance [7]. If we know which groups of attributes that are interacting with respect to the classes, it makes sense to perform variable subset selection from among these groups, thus keeping attribute interactions intact so that the classifier can exploit them. We demonstrate this in Sec. 6.

1.1 Contributions Our contributions are:

- We show how groups of interacting attributes are related to a factorisation of the conditional data distribution.
- We present and study the two problems of (i) assessing whether a particular grouping of attributes represents the class conditional structure of a data set using statistical hypothesis testing (Sec. 3) and (ii) automatically discovering the attribute grouping of highest granularity, with Automatic STRUCTURE Identification (ASTRID) (Sec. 4).
- A novel polynomial time clustering algorithm that relies on a certain monotonic, but in practice very intuitive score function (Sec. 4.1).
- We present an experimental verification and discuss several real-world data analysis scenarios that become possible through knowledge of the class-conditional joint data distribution (Secs. 5–7).

2 Related Work

This work has been motivated and influenced by the recently introduced **GoldenEye** algorithm [8, 9], as well as the results about classifiers and attribute interactions in [17]. [8] and the method introduced in this paper share some similarities, but ultimately address different problems: [8] reveals the structure of classifier function (even the structure imposed by overlearning), while the approach presented here reveals only the structure really present in the data *and* that can be modeled by training the classifier. As opposed to the heuristic used in [8], we approach the problem in a principled way using statistical significance testing.

The main result in this paper is a randomisation test for testing the hypothesis that an observed dataset has been sampled from a given class-conditional joint distribution with a particular factorised form. The problem considered here is closely related to the permutation test in [17], but instead of only considering a fully factorised

class conditional distribution (i.e., the one assumed by the naïve Bayes classifier), the test considered here is valid for any given factorisation. As a consequence our method can be used to reveal the attribute interactions present in the data in terms of the class-conditional joint distribution. Our second contribution is an algorithm for automatically determining the structure of the data in terms of the finding the factorisation with the highest granularity such that a classifier trained using the factorised data is indistinguishable from a classifier trained using the original data.

Moreover, various methods to study attribute interactions in general have been proposed, see, e.g., [6] for a review on the topic in data mining. Applications to feature selection has been studied by, e.g., [21, 22]. [20] investigated finding maximally dependent successively ordered attributes while [15] clustered correlated attributes into groups. [10] proposed a method for quantifying the degree of interaction and [11] consider factorising the joint data distribution and presented a method for testing the significance of the found attribute interactions (experiments limited to two and three-way interactions). Investigating the structure of the data in terms of factorising the joint distribution is also the topic of Bayesian network learning (e.g., [13]).

3 Verifying a Single Grouping

We first introduce the necessary notation, after which we present a hypothesis testing framework for studying attribute groupings.

3.1 Preliminaries Let X be an $n \times m$ data matrix, where $X(i, \cdot)$ denotes the i th row (item), $X(\cdot, j)$ the j th column (attribute) of X , and $X(\cdot, S)$ the columns of X given by S , where $S \subseteq [m] = \{1, \dots, m\}$, respectively. Let \mathcal{C} be a finite set of class labels and let C be an n -vector of class labels, such that $C(i)$ gives the class label for $X(i, \cdot)$. We denote a dataset D by the tuple $D = (X, C)$ and the set of all possible datasets by \mathcal{D} .

We denote by \mathcal{P} the set of disjoint partitions of $[m] = \{1, \dots, m\}$, where a partition $\mathcal{S} \in \mathcal{P}$ satisfies $\cup_{S \in \mathcal{S}} S = [m]$ and for all $S, S' \in \mathcal{S}$ either $S = S'$ or $S \cap S' = \emptyset$, respectively.

The dataset D follows a joint probability distribution

$$(3.1) \quad P(D) = \prod_{i \in [n]} \overbrace{P(X(i, \cdot) | C(i))}^{P(X|C)} P(C(i)),$$

where $P(X | C)$ is the *class-conditional distribution*. We consider a factorisation of $P(D)$ into class-conditional factors given by the grouping $\mathcal{S} \in \mathcal{P}$ and

write

$$(3.2) \quad P(D) = \prod_{i \in [n]} \prod_{S \in \mathcal{S}} \overbrace{P(X(i, S) | C(i))}^{\prod_{S \in \mathcal{S}} P(X(\cdot, S) | C)} P(C(i)).$$

Given an observed dataset $D_0 \in \mathcal{D}$, we want to investigate the structure of the data in terms of groupings $\mathcal{S} \in \mathcal{P}$ and a natural approach is to formulate a null hypothesis:

HYPOTHESIS 1. *The observed dataset D_0 has been sampled from a distribution given by Eq. (3.2) with the grouping given by $\mathcal{S} \in \mathcal{P}$.*

We now devise a framework to test this hypothesis.

3.2 Hypothesis Testing Framework Hypothesis 1 states that the dataset D_0 has been sampled from a distribution that follows the form given by Eq. (3.2) with the groups given by \mathcal{S} . This hypothesis can be evaluated empirically using a randomisation test, for which we need (i) a test statistic and (ii) the distribution of the test statistic under the null hypothesis. The value of the test statistic for the observed data is compared to the distribution of the test statistic under the null hypothesis. The outcome of the comparison is typically reported in the form of a p -value denoting the probability of obtaining a result at least as extreme as the observed one under the null hypothesis. Inference regarding the hypothesis is then made at a significance level α denoting the probability of a Type I error.

3.2.1 Test Statistic Assume for now that the test statistic yields a real number for each dataset in \mathcal{D} , i.e., $T : \mathcal{D} \mapsto \mathbb{R}$. The exact form of the test statistic used here is described in detail later in Sec. 3.3 after we have presented the general framework.

3.2.2 GoldenEye Permutation In this paper we sample datasets using the **GoldenEye** permutation, first described in [8]. The **GoldenEye** permutation is parametrized by a grouping $\mathcal{S} \in \mathcal{P}$ and creates a sample from the set of all datasets by permuting the columns of the original data within-class so that columns in the same group $S \in \mathcal{S}$ are permuted together.

More formally, a new dataset $D^{\mathcal{S}} = (X^{\mathcal{S}}, C)$ is created by permuting the data matrix of the dataset $D_0 = (X_0, C)$ at random. The permutation is defined by m bijective permutation functions $\pi_j : [n] \mapsto [n]$ sampled uniformly at random from the set of allowed permutations functions. The new data matrix is then given by $X^{\mathcal{S}}(i, j) = X_0(\pi_j(i), j)$. The allowed permutation functions satisfy the following conditions for all $i \in [n]$, $j, j' \in [m]$, and $S \in \mathcal{S}$:

1. permutations are within-a class, i.e., $C(i) = C(\pi_j(i))$, and
2. items within a group are permuted together, i.e., $j \in S \wedge j' \in S \implies \pi_j(i) = \pi_{j'}(i)$.

Let $\mathcal{D}_S \subseteq \mathcal{D}$ be the set of datasets that can be generated using the **GoldenEye** permutation with the grouping \mathcal{S} . We make the following two observations.

LEMMA 3.1. *Each invocation of the **GoldenEye** permutation produces each of the datasets in \mathcal{D}_S with uniform probability.*

LEMMA 3.2. *The datasets in \mathcal{D}_S have equal probability under the distribution of Eq. (3.2), parametrized by \mathcal{S} .*

We omit the proofs for brevity, which follow directly from the definitions. It follows from Lemmas 3.1 and 3.2 that the data sampled using **GoldenEye** obeys the distribution of Eq. (3.2) in \mathcal{D}_S and, hence, an empirical p -value defined as follows is a valid, i.e., it is stochastically larger than the unit distribution in $[0, 1]$ under the null hypothesis that the data originates from the distribution given by Eq. (3.2),

$$(3.3) \quad p_S = \frac{1 + \sum_{i=1}^R I[T(D_i^S) \geq T(D_0)]}{1 + R},$$

where $I[o]$ is the indicator function which equals unity if o is true and zero otherwise, $T : \mathcal{D} \mapsto \mathbb{R}$ is the test statistic, D_0 is the original observed dataset, and $D_i^S \in \mathcal{D}_S$, where $i \in [R]$, are R samples created by the **GoldenEye** permutation parametrized by \mathcal{S} . If we here obtain $p_S < \alpha$, we can reject the null hypothesis that the dataset D_0 could have been generated by the distribution given by Eq. (3.2) with a significance level α . We next specify the exact form of the test statistic T .

3.3 Hypothesis Testing Using Classifiers The above described framework is valid for any test statistic T , but a poor choice of T could result to unnecessarily large number of Type II errors, i.e., failure to detect an interaction of attributes. The test statistic should reflect how well \mathcal{S} captures the structure of the data. As discussed in Sec. 1, it is reasonable to assume that high-performing classifiers internally model the class-conditional joint distribution of the attributes (Eq. (3.2)). A classifier is a function that tries to predict classes given a row of data matrix. A classifier is typically generated using a training data set containing both the data matrix and the class vector $D = (X, C)$. We denote a classifier trained using the data set D by $f_D : \mathcal{X} \mapsto \mathcal{C}$, where \mathcal{X} denotes the set of all possible

rows of the data matrix. Further assume that we have a separate independent test data set from the same distribution as D_0 , denoted by $D_{\text{test}} = (X_{\text{test}}, C_{\text{test}})$. With these components, we define the test statistic as

DEFINITION 1. *Test Statistic Given the above definitions, the test statistic for a dataset $D \in \mathcal{D}$ is given by*

$$(3.4) \quad T(D) = \frac{1}{n_{\text{test}}} \sum_{i=1}^{n_{\text{test}}} I[f_D(X_{\text{test}}(i, \cdot)) = C_{\text{test}}(i)],$$

where n_{test} is the number of items in the test data set.

We have here chosen, for simplicity to use accuracy, but other performance metrics could be used as well, e.g., the F_1 measure.

Finally, we cast the above presented hypothesis testing procedure in the form of a problem:

PROBLEM 1. *Given an observed dataset $D_0 \in \mathcal{D}$, a grouping \mathcal{S} , and a classifier f , determine at a level $\alpha \in [0, 1]$ if D_0 has been sampled from a distribution given by Eq. (3.2) such that the factors are given by \mathcal{S} .*

To solve Prob. 1 we proceed as follows: (i) use the test statistic of Definition 1 with the classifier f , (ii) determine the distribution of the test statistic under the null hypothesis from datasets generated from the observed dataset using the **GoldenEye** permutation parametrised by the grouping \mathcal{S} , and finally (iii) compute the p -value in Eq. (3.3), after which we evaluate the null hypothesis at the significance level α . If we find $p_S \geq \alpha$ we conclude that the null hypothesis cannot be rejected and hence we cannot rule out that D_0 originates from a distribution given by Eq. (3.2) with groups given by \mathcal{S} .

4 Automatically Finding Groupings (ASTRID)

The above described method allows us to test whether a particular grouping \mathcal{S} describes the structure of the data in terms of the factorisation in Eq. (3.2). The following problem is a natural extension of the above discussion:

PROBLEM 2. *Given an observed dataset D_0 and a classifier f , find the grouping \mathcal{S} of cardinality k such that the accuracy is maximized.*

Instead of specifying the cardinality k in advance, we can also use the confidence level α as follows: after finding the solution to Prob. 2, we compute the p -value using Eq. (3.3) for solutions of all cardinalities $k = 2, \dots, m$. The desired solution is given by the highest-cardinality grouping satisfying $p_S \geq \alpha$.

Interestingly, Prob. 2 can be viewed as an instance of a generic clustering problem:

PROBLEM 3. Given integers k and m and a reward function $\hat{T} : \mathcal{P} \mapsto \mathbb{R}$, where \mathcal{P} is the set of all partitions of $[m]$, find a partition $\mathcal{S} \in \mathcal{P}$ of size $|\mathcal{S}| = k$ such that the reward $\hat{T}(\mathcal{S})$ is maximized.

Prob. 2 reduces to Prob. 3 if we use the expected accuracy T defined in Eq. (3.4) as the the reward function \hat{T} in the clustering problem;

$$(4.5) \quad \hat{T}(\mathcal{S}) = \frac{1}{R'} \sum_{i=1}^{R'} T(D_i^{\mathcal{S}}),$$

where $D_i^{\mathcal{S}}$, $i \in [R']$, is a random dataset produced by the **GoldenEye** permutation parametrized by \mathcal{S} . We provide a polynomial time heuristic algorithm to solve Prob. 3. The algorithm yields the exact solution to the problem if the accuracy is monotonic (Def. 2 and Theorem 4.1). The clustering algorithm and its properties are described in more detail later in Sec. 4.1.

The monotonicity of the accuracy means that if \mathcal{S}_0 is the solution to Prob. 2, then the accuracy $\hat{T}(\mathcal{S})$ is reduced if any group in \mathcal{S}_0 has been broken in \mathcal{S} . More specifically, let $\mathcal{S} = \{S_1, \dots, S_k\}$ be a partition of $[m]$, and let $\mathcal{S}' = \{S_{1a}, S_{1b}, S_2, \dots, S_k\}$, where $S_1 = S_{1a} \cup S_{1b}$ and $S_{1a} \cap S_{1b} = \emptyset$. Assume there exist a group $Q \in \mathcal{S}_0$ such that $Q \cap S_1 = Q$, but $Q \cap S_{1a} \neq Q$ and $Q \cap S_{1b} \neq Q$, i.e., splitting S_1 into S_{1a} and S_{1b} has broken at least the group Q in \mathcal{S}_0 . The monotonicity of the accuracy implies here that $\hat{T}(\mathcal{S}') < \hat{T}(\mathcal{S})$, i.e., breaking the interactions in Q means that the classifier cannot utilize them fully, which is expected to reduce classification accuracy. If the monotonicity is preserved to a sufficient accuracy, then we expect that the clustering algorithm can efficiently and accurately provide a solution to Prob. 2.

Solving Prob. 2 for all k requires evaluation of the accuracy $\hat{T}(\mathcal{S})$ for $\mathcal{O}(m^2)$ different values of \mathcal{S} . If we want to further find the highest-cardinality grouping satisfying $p_{\mathcal{S}} \geq \alpha$ then $\mathcal{O}(m)$ p -value computations are additionally needed which, however, does not increase the computational complexity compared to just solving Prob. 2.

4.1 Clustering Problem As discussed in Sec. 4, to find an optimal grouping we need to solve a generic clustering problem (Prob. 3).

If no assumptions of the reward function are made then verifying that a given partition is a solution to Prob. 3 requires evaluation of $\hat{T}(\mathcal{S})$ for all partitions $\mathcal{S} \in \mathcal{P}$ of size k . In order to provide a polynomial time algorithm to solve the problem we hence need to make some assumptions regarding the form of the reward function. We can indeed devise an efficient and exact algorithm if \hat{T} behaves consistently in the sense that the

reward function decreases if any of the clusters in the (a priori unknown) solution to Prob. 3 are broken. We call this property *monotonicity* and define it formally as follows.

DEFINITION 2. For given integers k and m and a reward function \hat{T} , as defined above, let \mathcal{S}_0 be the solution to Prob. 3. The reward function \hat{T} is monotonic iff all $\mathcal{S}, \mathcal{S}' \in \mathcal{P}$ satisfying $F(\mathcal{S}) \subset F(\mathcal{S}')$ also satisfy $\hat{T}(\mathcal{S}) < \hat{T}(\mathcal{S}')$, where we have used $F(\mathcal{S}) = \{X \in \mathcal{S}_0 \mid \exists Y \in \mathcal{S}. X \subseteq Y\}$.¹

We propose a heuristic clustering algorithm, described in Alg. 1, for solving Prob. 3. The algorithm first sorts the attributes by iteratively moving attributes to singleton clusters such that the accuracy is maximized at each step (lines 1–6). After this the algorithm finds the k -segmentation of the ordered set of attributes (lines 7–9) corresponding to the solution of Prob. 3 (lines 10–11). Even though the algorithm is heuristic in the general case, it provides an exact solution if the reward function is monotonic.

THEOREM 4.1. Alg. 1 solves Prob. 3 exactly if the reward function \hat{T} is monotonic.

Proof. Denote by \mathcal{S}_0 the solution to Prob. 3. In the sorting phase of Alg. 1 (lines 1–6) the attributes are ordered into a vector a_1, \dots, a_m so that all clusters $S \in \mathcal{S}_0$ appear in a continuous segment. This follows directly from the monotonicity: consider an iteration of this loop (lines 3–5). Let $S \in \mathcal{S}_0$ be the cluster that contains j obtained in line 3, i.e., $j \in S$. If we have not yet processed all attributes in S , i.e., if after line 4 $S \cap C \neq \emptyset$, then in the next iteration of the loop we must choose a value of j from $S \cap C$, because by monotonicity choosing j not in $S \cap C$ would result in a lower reward than choosing j from $S \cap C$.

Therefore, because the vector a_1, \dots, a_m contains the clusters in \mathcal{S}_0 in continuous segments, the problem reduces to finding $k - 1$ segment boundaries that split the vector into k segments, with each segment corresponding to a cluster in \mathcal{S}_0 . We can make an observation that if we split the attributes in $[m]$ into two clusters $L \subseteq [m]$ and $[m] \setminus L$, then the reward $\hat{T}(\{L, [m] \setminus L\})$ is larger if the split into two clusters does not break clusters in \mathcal{S}_0 . In other words, $\hat{T}(\{L, [m] \setminus L\}) > \hat{T}(\{L', [m] \setminus L'\})$ if there is a subset of clusters $R \subseteq \mathcal{S}_0$ such that $L = \cup_{X \in R} X$ and there is no subset of clusters $R' \subseteq \mathcal{S}_0$ such that $L' = \cup_{X \in R'} X$. Therefore, it suffices to compute the costs of all segmentations into two (line 7), and pick the $k - 1$ segment

¹If for example $\mathcal{S}_0 = \{\{1, 2\}, \{3, 4\}\}$ then $F(\{\{1, 2, 3\}, \{4\}\}) = \{\{1, 2\}\}$, i.e., the function $F(\mathcal{S})$ returns members of \mathcal{S}_0 that are unbroken in \mathcal{S} .

```

input :  $k, m$ , and  $\hat{T}$  as defined in Prob. 3
output: partition  $\mathcal{S}$  of size  $k$  maximizing  $\hat{T}(\mathcal{S})$ 
/* Sorting */
1 Let  $B \leftarrow \emptyset, C \leftarrow [m]$ ;
2 for  $i = 1$  to  $m$  do
3   Let  $j \leftarrow$ 
    $\arg \max_{j \in C} \hat{T}(\{\{C \setminus \{j\}\} \cup \bigcup_{l \in B \cup \{j\}} \{\{l\}\}\});$ 
4   Let  $B \leftarrow B \cup \{j\}, C \leftarrow C \setminus \{j\}$ ;
5   Let  $a_i \leftarrow j$ ;
6 end
/* Grouping */
7 Let  $t_i \leftarrow \hat{T}(\{\{a_1, \dots, a_i\}, \{a_{i+1}, \dots, a_m\}\})$  for all
 $i \in [m-1]$ ;
8 Let  $i_0 \leftarrow 0, i_k \leftarrow m$ ;
9 Let  $i_1 < \dots < i_{k-1}$  be such that  $\{t_{i_1}, \dots, t_{i_{k-1}}\}$ 
are the  $k-1$  largest values of  $\{t_1, \dots, t_{m-1}\}$ ;
10 Let  $S_j \leftarrow \{a_{i_{j-1}+1}, \dots, a_{i_j}\}$  for all  $j \in [k]$ ;
11 Let  $\mathcal{S} \leftarrow \{S_1, \dots, S_k\}$ ;
12 return  $\mathcal{S}$ 

```

Algorithm 1: The clustering algorithm.

boundaries corresponding to the highest rewards (line 9). The resulting segments, defined in line 10, must then correspond to the clusters in \mathcal{S}_0 . \square

Notice that Alg. 1 can also be used to efficiently give the clustering for all values of k since lines 1–7 are common for all values of k and only lines 8–11 need to be re-run for different values of k . The clustering algorithm therefore requires only $\mathcal{O}(m^2)$ evaluations of the reward function to find clusterings for all values of $k \in [m]$.

Moreover, usually clustering cost functions are defined in terms of distances between cluster centroids or data points. In our case we do not, however, have any well-defined distances between data points and, hence, normal clustering algorithms are not applicable. Instead, we have to find the correct grouping based on the value of the reward (cost) function alone, which makes the problem more challenging. However, the monotonicity assumption of Def. 2 allows us to in fact find optimal solutions in polynomial time. To the best of our knowledge, this particular approach to finding clusterings has not been considered previously, and we think it may have applications also in other contexts.

5 Experiments

Experimental setup We evaluate the method proposed in this paper empirically addressing three case examples demonstrating the utility of our method. More specifically, we show that the proposed ASTRID method allows us to (1) identify attribute interactions modelled by the classifier in a dataset, (2) generate (anonymised)

surrogate datasets with the same conditional distribution as an original dataset, and (3) fuse datasets.

In the experiments we use the synthetic dataset (Fig. 1) and 9 datasets from the UCI machine learning repository [2]². All experiments were run in R [18] and the method is released as the ASTRID R-package. The ASTRID R-package and the source code for the experiments are available for download³.

All statistical significance testing in the experiments is conducted at the $\alpha = 0.05$ level. We use a value of $R \geq 250$ ⁴ and $R' = 100$ in Eqs. (3.3) and (4.5), respectively. In all experiments the dataset was randomly split as follows: 50% for training (D_0) and the rest for testing data sets (D_{test} , see Eq. (3.4)): 25% for computing of \hat{T} (Eq. (4.5)), and 25% to find the highest-cardinality grouping satisfying $p_{\mathcal{S}} \geq \alpha$ from among the results of the ASTRID method.

Classifiers Classifier choice is important. The SVM and random forest classifiers are among the best-performing classifiers [5] and we hence use these (SVM with RBF kernel from the `e1071` R-package [16] and random forest from the `randomForest` [14] package). We also show some examples with a naïve Bayes classifier (from `e1071`). All classifiers were used at their default settings.

Datasets The properties of the datasets are summarised in Table 1, also showing the computation time for the SVM and RF. The computation time of the ASTRID method depends both on the properties of the dataset such as the number of attributes and instances, and on the used classifier. The UCI datasets were chosen to have at least 600 items and so that the SVM and random forest classifiers achieve reasonably good accuracy at default settings, since the goal here is to demonstrate the applicability of the method rather than optimise classifier performance. Rows with missing values and constant-value columns were removed from the UCI datasets.

6 Results

We now present empirical results obtained using the ASTRID method and demonstrate the usefulness and applicability of the method in three real-world contexts.

6.1 Finding Attribute Interactions The results are presented as tables where the columns represent

²Datasets obtained from <http://www.cs.waikato.ac.nz/ml/weka/datasets.html>

³<https://github.com/bwrc/astrid-r>

⁴At least 250 samples are used in an early stopping scheme for p -value calculation.

Table 1: The datasets used in the experiments (2–10 from UCI). Columns as follows: Number of items (Ni) after removal of rows with missing values, number of classes (Nc) after removal of constant-value columns, number of attributes (Na). MCP is major class proportion. T_{SVM} and T_{RF} give the calculation in minutes of the ASTRID method for the SVM and random forest, respectively.

n	Dataset	Ni	Nc	Na	MCP	T_{SVM}	T_{RF}
1	synthetic	1000	2	4	0.50	0.1	0.4
2	balance-scale	625	3	4	0.46	0.7	4.0
3	diabetes	768	2	8	0.65	2.9	13.3
4	vowel	990	11	13	0.09	15.4	226.1
5	credit-a	653	2	15	0.55	9.3	40.0
6	segment	2310	7	18	0.14	39.9	110.2
7	vehicle	846	4	18	0.26	16.5	71.4
8	mushroom	5644	2	21	0.62	107.0	134.6
9	soybean	682	19	35	0.13	79.7	181.1
10	kr-vs-kp	3196	2	36	0.52	186.0	200.7

Table 2: The synthetic dataset.

SVM						Random forest						naïve Bayes							
k	acc	p	a ³	a ⁴	a ¹	k	acc	p	a ³	a ⁴	a ¹	a ²	k	acc	p	a ¹	a ²	a ³	a ⁴
2	0.89	0.72	(A)	(B)	(B)	2	0.91	0.82	(A)	(B)	(B)	2	0.73	1.00	(A)	(B)	(B)		
3	0.88	0.78	(A)	(B)	(C)	3	0.91	0.83	(A)	(B)	(C)	3	0.73	1.00	(A)	(B)	(C)		
4	0.74	0.00	(A)	(B)	(C)	4	0.74	0.00	(A)	(B)	(C)	4	0.73	1.00	(A)	(B)	(C)		
{1, 2}, {3}, {4}						{1, 2}, {3}, {4}						{1}, {2}, {3}, {4}							

attributes, sorted in the order in which the attributes were detached in the sorting step. Each row represents a grouping. Attributes belonging to the same group are marked with the same letter, i.e., attributes marked with the same letter are interacting. The highest-cardinality grouping for which $p \geq 0.05$ is marked with green and this grouping is also shown below the table.

Table 2 shows the results for the synthetic dataset. The groupings of size $k = 2$ and $k = 3$ are valid ($p \geq 0.05$) for SVM and random forest. For naïve Bayes also $k = 4$ is valid. The SVM and random forest classifiers both identify the correct interaction structure of the dataset. The naïve Bayes classifier always assumes that each attribute is independent and

Table 3: Grouping of the **credit-a** dataset using SVM.

k	acc	p	A14	A7	A13	A1	A6	A12	A15	A8	A9	A2	A3	A5	A4	A11	A10
2	0.86	0.53	(A)	(A)	(A)	(B)	(B)	(B)	(B)	(B)	(B)	(B)	(B)	(B)	(B)	(B)	(B)
3	0.86	0.43	(A)	(A)	(A)	(B)	(B)	(C)	(C)	(C)	(C)	(C)	(C)	(C)	(C)	(C)	(C)
4	0.86	0.37	(A)	(A)	(A)	(B)	(C)	(D)	(D)	(D)	(D)	(D)	(D)	(D)	(D)	(D)	(D)
5	0.85	0.46	(A)	(A)	(A)	(B)	(C)	(D)	(E)	(E)	(E)	(E)	(E)	(E)	(E)	(E)	(E)
6	0.85	0.49	(A)	(A)	(B)	(C)	(D)	(E)	(F)	(F)	(F)	(F)	(F)	(F)	(F)	(F)	(F)
7	0.85	0.45	(A)	(A)	(B)	(C)	(D)	(E)	(F)	(G)	(G)	(G)	(G)	(G)	(G)	(G)	(G)
8	0.85	0.44	(A)	(B)	(C)	(D)	(E)	(F)	(G)	(H)	(H)	(H)	(H)	(H)	(H)	(H)	(H)
9	0.85	0.25	(A)	(B)	(C)	(D)	(E)	(F)	(G)	(H)	(I)	(I)	(I)	(I)	(I)	(I)	(I)
10	0.84	0.10	(A)	(B)	(C)	(D)	(E)	(F)	(G)	(H)	(I)	(I)	(I)	(J)	(J)	(J)	(J)
11	0.83	0.04	(A)	(B)	(C)	(D)	(E)	(F)	(G)	(H)	(I)	(J)	(K)	(K)	(K)	(K)	(K)
12	0.83	0.03	(A)	(B)	(C)	(D)	(E)	(F)	(G)	(H)	(I)	(J)	(K)	(L)	(L)	(L)	(L)
13	0.82	0.04	(A)	(B)	(C)	(D)	(E)	(F)	(G)	(H)	(I)	(J)	(K)	(L)	(L)	(M)	(M)
14	0.82	0.04	(A)	(B)	(C)	(D)	(E)	(F)	(G)	(H)	(I)	(J)	(K)	(L)	(L)	(N)	(N)
15	0.82	0.03	(A)	(B)	(C)	(D)	(E)	(F)	(G)	(H)	(I)	(J)	(K)	(L)	(M)	(N)	(O)

$S_{10} = \{\{14\}, \{7\}, \{13\}, \{1\}, \{6\}, \{12\}, \{15\}, \{8\}, \{9, 2, 3\}, \{5, 4, 11, 10\}\}$

Table 4: Groupings for the synthetic and UCI datasets using SVM and random forest. The columns are as follows. The number of attributes in the dataset (N), the size of the grouping (k), the size of the largest (N_1) and second-largest (N_2) groups, and the significance of the grouping (p). The other columns are: baseline accuracy when the classifier is trained with unshuffled data (a_0) and with data shuffled using the found grouping (a), the range of the accuracy (a_{range}) and its standard deviation (a_{sd}). The column p_{OG} is the p -value corresponding to Test 2 in [17].

SVM										
Dataset	N	k	N_1	N_2	p	a_0	a	a_{range}	a_{sd}	p_{OG}
bal-sc.	4	3	2	1	0.14	0.89	0.86	[0.79, 0.92]	0.02	0.03
credit-a	15	10	4	3	0.10	0.87	0.85	[0.83, 0.88]	0.01	0.04
diabetes	8	8	1	1	0.59	0.71	0.71	[0.67, 0.75]	0.01	0.59
kr-vs-kp	36	33	4	1	1.00	0.92	0.92	[0.92, 0.92]	0.00	0.00
mushr.	21	15	7	1	0.06	1.00	0.99	[0.99, 1.00]	0.00	0.00
segment	18	2	16	2	0.07	0.95	0.94	[0.93, 0.95]	0.00	0.00
soybean	35	35	1	1	0.35	0.84	0.84	[0.81, 0.86]	0.01	0.26
vehicle	18	5	14	1	0.07	0.77	0.74	[0.70, 0.79]	0.02	0.00
vowel	13	3	11	1	0.09	0.81	0.78	[0.72, 0.82]	0.01	0.00

random forest										
Dataset	N	k	N_1	N_2	p	a_0	a	a_{range}	a_{sd}	p_{OG}
bal-sc.	4	3	2	1	0.11	0.82	0.78	[0.69, 0.86]	0.03	0.02
credit-a	15	15	1	1	0.39	0.88	0.86	[0.82, 0.90]	0.01	0.19
diabetes	8	8	1	1	0.90	0.70	0.72	[0.68, 0.76]	0.01	0.89
kr-vs-kp	36	13	24	1	0.08	0.98	0.98	[0.97, 0.98]	0.00	0.00
mushr.	21	14	8	1	0.19	1.00	1.00	[0.99, 1.00]	0.00	0.00
segment	18	5	14	1	0.15	0.99	0.98	[0.97, 0.99]	0.00	0.00
soybean	35	23	11	3	0.07	0.96	0.95	[0.94, 0.96]	0.00	0.00
vehicle	18	5	12	2	0.06	0.75	0.73	[0.70, 0.77]	0.01	0.00
vowel	13	3	11	1	0.37	0.92	0.91	[0.89, 0.93]	0.01	0.00

all groupings are equally valid as no interactions are utilised. These results mean that an SVM trained on the synthetic dataset **GoldenEye**-permuted using $S = \{\{1, 2\}, \{3\}, \{4\}\}$ is indistinguishable (at the 5% level) from an SVM trained on the original synthetic dataset. We hence find the factorised form of the joint distribution of the data.

Table 3 shows the valid groupings of the **credit-a** dataset using SVM. This dataset contains many attributes not used by the SVM and the dataset can be permuted to a large extent without impacting classifier performance. The grouping with $k = 10$ contains two larger groups of interacting attribute, marked with I and J , respectively, while the other attributes are singletons. In variable subset selection, as discussed in Example 3 in Sec. 1, it is meaningful to pick attributes based on their interaction. Here it would hence be meaningful to consider the attributes in the I and J groups for $k = 10$ when selecting features, since the attributes in these groups are interacting.

The groupings for the UCI datasets are summarised in Table 4, showing the properties in terms of the

number of attributes used by the classifiers and the number of groups and singletons together with statistics describing the accuracy of the classifier using the found grouping. The groupings for the SVM and random forest are quite similar in terms of the size of the found grouping (k in Table 4) and the number of singletons, although the structures utilised by the classifiers is somewhat different.

To compare our results with those of [17], investigating whether a classifier utilises attribute interactions, we computed the p -value for their Test 2 (denoted p_{OG} in Table 4), which is equivalent to our Problem 1 with an all-singleton grouping. $p_{\text{OG}} \geq 0.05$ indicates that the classifier does not utilise attribute interactions in the dataset. This occurs for `diabetes` and `soybean` for SVM and for `diabetes` and `credit-a` for random forest. This is in line with our findings, since for these datasets k equals N in Table 4 and no interactions are utilised.

6.2 Anonymisation and Surrogate Data Data anonymisation (e.g., [1, 3, 19]) and synthetic data generation are related processes with the goal of generating data that shares properties with the original data so that the anonymised data can be used in place of the original data. Anonymisation can be done in a principled way if we know the attribute interactions in the dataset. One method of data anonymisation is based on shuffling the data, which we employ here. As a measure of anonymity we compute the proportion P_{anon} of identical items in the original dataset still present after shuffling.

We consider the UCI `credit-a` dataset. It contains information on credit card applications and has 15 attributes and two classes: positive or negative decision. We here use the grouping for $k = 10$ in Table 3, shown as \mathcal{S}_{10} below the table.

The anonymised dataset is obtained from the original dataset by permuting it with \mathcal{S}_{10} . The accuracy using original, unshuffled data is 0.87 and when the classifier is trained with anonymised data the accuracy is 0.85 (average of 100 repetitions). To test the quality of the anonymisation, we calculated P_{anon} for 100 repetitions. The original training dataset has 327 unique rows and we obtain an average $P_{\text{anon}} = 0$, i.e., the anonymisation is very efficient. Shuffling the data using the `GoldenEye` permutation with \mathcal{S}_{10} yields new surrogate datasets with the same class-conditional distribution as the original dataset.

6.3 Efficient Data Fusion and Collection Assume that we have a small dataset and have computed its structure using ASTRID. Further assume that we

want to collect more data to be used as a training dataset for a classifier, and also assume that we can assign class labels to data items from an external source.

An interesting implication of the method presented here is that it can be used for efficient data fusion and collection. E.g., for the `credit-a` dataset a data collection task can be split up into the sets given by \mathcal{S}_{10} in Table 3. We may collect a new dataset by letting one survey participant respond to only the attributes in some of the sets in \mathcal{S}_{10} and tag the answers with the class, which for `credit-a` is the known credit decision. Independently collected answers regarding different sets of attributes can then be fused based on the class label and used to train a new classifier. This works, because if the distribution of the data obeys the class-conditional form of Eq. (3.2), then it should not matter in training of a classifier if different attribute groups on a row of the data matrix are in fact collected from separate persons in the same class.

Collecting survey data is often costly and time-consuming as a large number of people must answer a large number of questions. Subdividing a survey therefore reduces answering time and costs. Using the above described method we may hence speed up data collection for a training dataset for a classifier.

In this manner new training data for a classifier can be collected more efficiently, since it can be partitioned into independent sets based on grouping of an initial dataset using ASTRID.

6.4 Summary of Results ASTRID (based on the presented hypothesis testing framework) finds the most granular grouping \mathcal{S} such that a classifier trained with data permuted by \mathcal{S} is indistinguishable in terms of accuracy from a classifier trained using the unpermuted data. The joint distribution can be factorised using \mathcal{S} and this structure can be leveraged as shown in the examples here.

7 Discussion and Conclusion

We present an efficient framework for testing the hypothesis that the class-conditional joint distribution of a dataset follows a specific factorised form. The factorised joint data distribution tells us what attribute interactions are used by a classifier and hence also what the structure of the data is. Knowledge of the joint distribution is important for data exploration and can be used in solving several real-world data analysis problems, in this paper exemplified by showing the utility in (i) finding attribute interactions (applications to variable selection), (ii) data anonymisation and generation of surrogate data, and (iii) data fusion. Our empirical investigation shows that often many interactions in the

UCI datasets considered here can be broken substantially without affecting classification performance.

The framework is realised in the ASTRID method which is made available as the ASTRID R-package⁵. It allows the factorisation of highest granularity to be automatically be found. Both the framework and ASTRID make no assumptions regarding the data distribution or the used classifier and hence have high generic applicability to different datasets and problems. The methods presented here build upon and extend ideas discussed in [8, 17]. However, unlike the groupings found in, e.g., [8], the framework presented here provides a statistical guarantees.

The relationship between statistical significance and practical relevance should be considered. E.g., although a negligible drop in performance due to factorisation of the joint distribution may be statistically significant, this decrease lacks practical relevance. Compare, e.g., the accuracies for $k = 2$ and $k = 15$ in Table 3.

A potential direction of future research is to extend the method to regression problems, requiring a redefinition of the permutation scheme so that the resulting datasets are samples from a factorised distribution conditioned on the dependent variable of the regression model. Also techniques to speed up the p -value computations by reducing the number of random samples required are of interest. Furthermore, the clustering problem (Sect. 4.1) and the associated monotonicity definition (Def. 2) may have applications also to other problems.

Acknowledgements This work was supported by Academy of Finland (decision 288814) and Tekes (Revolution of Knowledge Work project).

References

- [1] R. Agrawal and R. Srikant. Privacy-preserving data mining. In *SIGMOD 2000*, pages 439–450, 2000.
- [2] K. Bache and M. Lichman. UCI machine learning repository, 2014.
- [3] R. J. Bayardo Jr. and R. Srikant. Technological solutions for protecting privacy. *IEEE Computer*, 36(9):115–118, 2003.
- [4] C. Bucilă, R. Caruana, and A. Niculescu-Mizil. Model compression. In *KDD 2006*, pages 535–541, 2006.
- [5] M. Fernández-Delgado, E. Cernadas, S. Barro, and D. Amorim. Do we need hundreds of classifiers to solve real world classification problems? *Journal of Machine Learning Research*, 15:3133–3181, 2014.
- [6] A. A. Freitas. Understanding the crucial role of attribute interaction in data mining. *Artificial Intelligence Review*, 16(3):177–199, 2001.
- [7] I. Guyon and A. Elisseeff. An introduction to variable and feature selection. *Journal of machine learning research*, 3(Mar):1157–1182, 2003.
- [8] A. Henelius, K. Puolamäki, H. Boström, L. Asker, and P. Papapetrou. A peek into the black box: exploring classifiers by randomization. *Data mining and knowledge discovery*, 28(5-6):1503–1529, 2014.
- [9] A. Henelius, K. Puolamäki, I. Karlsson, J. Zhao, L. Asker, H. Boström, and P. Papapetrou. Goldeneye++: A closer look into the black box. In *Statistical Learning and Data Sciences*, pages 96–105. Springer, 2015.
- [10] A. Jakulin and I. Bratko. Analyzing attribute dependencies. In *PKDD 2003*, pages 229–240. Springer, 2003.
- [11] A. Jakulin and I. Bratko. Testing the significance of attribute interactions. In *ICML 2004*, 2004.
- [12] A. Jakulin, I. Bratko, D. Smrke, J. Demšar, and B. Zupan. Attribute interactions in medical data analysis. In *Conference on Artificial Intelligence in Medicine in Europe*, pages 229–238. Springer, 2003.
- [13] T. J. Koski and J. M. Noble. A review of bayesian networks and structure learning. *Annales Societatis Mathematicae Polonae. Series 3: Mathematica Applicanda*, 40(1):53–103, 2012.
- [14] A. Liaw and M. Wiener. Classification and regression by randomforest. *R News*, 2(3):18–22, 2002.
- [15] M. Mampaey and J. Vreeken. Summarizing categorical data by clustering attributes. *Data Mining and Knowledge Discovery*, 26(1):130–173, 2013.
- [16] D. Meyer, E. Dimitriadou, K. Hornik, A. Weingessel, and F. Leisch. *e1071: Misc Functions of the Department of Statistics (e1071)*, TU Wien, 2014. R package version 1.6-4.
- [17] M. Ojala and G. C. Garriga. Permutation tests for studying classifier performance. *Journal of Machine Learning Research*, 11:1833–1863, 2010.
- [18] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2015.
- [19] L. Sweeney. k-anonymity: A model for protecting privacy. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 10(05):557–570, 2002.
- [20] N. Tatti. Are your items in order. In *SDM 2011*, pages 414–425. SIAM, 2011.
- [21] Z. Zhao and H. Liu. Searching for interacting features. In *IJCAI 2007*, pages 1156–1161, 2007.
- [22] Z. Zhao and H. Liu. Searching for interacting features in subset selection. *Intell. Data Anal.*, 13(2):207–228, 2009.

⁵<https://github.com/bwrc/astrid-r>